CS 686: Special Topics in Big Data

Data Sources & Network Design

Lecture 2
8/25/17

Today's Agenda

- **Q&A from the previous class**
- Defining big data
- Dataset sources
- Distributed network design

Q&A From the Previous Class

- A good book to brush up on distributed systems concepts:
  - *Distributed Systems: Principles and Paradigms* by Andrew S. Tanenbaum and Maarten van Steen
  - Not required
- We'll focus on relevant conference/journal papers for our readings
  - Two papers are available on the schedule page

Q&A From the Previous Class

- Project 1 will be implemented in Java
  - The majority of modern distributed systems are written in Java, or target the JVM
- Projects 2 and 3 will give you some flexibility in the language department

Today's Agenda

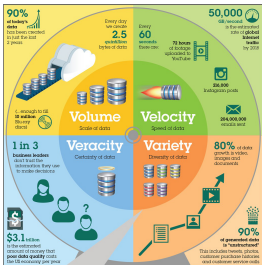
- Q&A from the previous class
- **Defining big data**
- Dataset sources
- Distributed network design

Defining Big Data

- In the last class, we talked about what "big data" really means
- The main takeaway is: it's hard to define!
- We can view it from different perspectives:
  - Systems, Machine Learning, Data Streams
  - The raw size of the data, what format it's in, processing required, etc…
- Let's look at this a little more in depth

The Four V's of Big Data

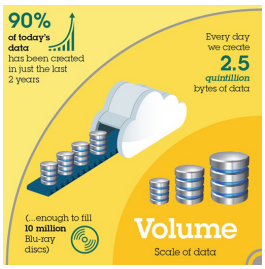
- Diving a bit deeper, we can divide the field up into four V's:
  - **V**olume
  - **V**elocity
  - **V**ariety
  - **V**eracity
- Some folks also include a fifth V – **V**alue.



Source: IBM. *Extracting business value from the 4 V's of big data.* http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data

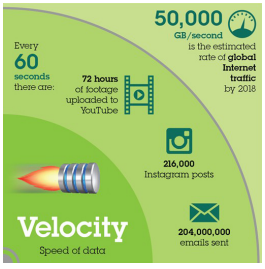
Volume

- Probably the easiest to understand!
- Voluminous datasets are managed by **distributed storage systems**
  - Google File System (GFS)
  - Amazon Dynamo
  - Apache Cassandra



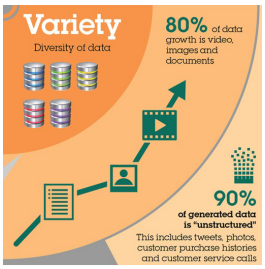
Velocity

- It's not enough to just be able to store a lot of data
- What happens if data comes in faster than our disks can write it?
- **Stream/Event Processing Systems**
  - Storm, Heron (Twitter)
  - Aurora, Samza



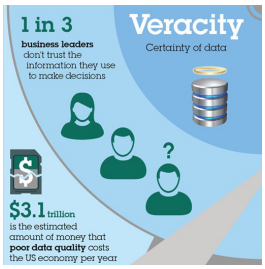
Variety

- Recent pushes towards "noSQL" and "newSQL" due to variety
  - Unstructured data is particularly relevant
  - Object-relational impedance mismatch
- Many systems specialize for a particular data type



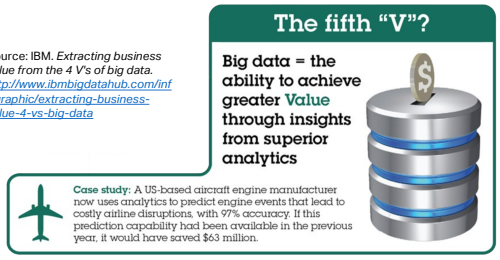
Veracity

- Does the data represent the full story?
- As always, correlation does not imply causation
- Datasets often require cleanup, have missing records, etc.
  - How do you handle these?



Value

Source: IBM. *Extracting business value from the 4 V's of big data.* http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data



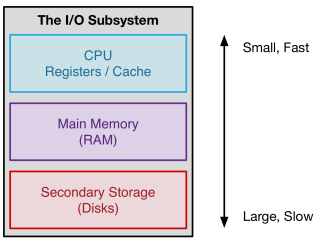
...So what makes data "big?"

- There isn't really a cutoff:
  - ~~Your data is larger than 1 petabyte, so it's big!~~
  - ~~I have 5 billion files, so that's big, right?~~
- Big data is more about the **scale**
  - In order to achieve your goals, you have to operate at a large scale
  - You may only have 1 TB of data, but it takes 72 hours to process on a single machine
  - You could have 10 PB of data that you process quickly, but you're limited by the I/O subsystem

The I/O Subsystem [1/3]

- One of the constant points of contention in Big Data is managing the speed differential of the memory hierarchy
- Modern computing systems have a variety of storage options with varying levels of speed and capacity

The I/O Subsystem [2/3]



The I/O Subsystem [3/3]

- Most big data lives on secondary storage
- The challenge? Getting it off of the disks and into cache, memory, or even the network
- Thought experiment:
  - Memory accesses are measured in nanoseconds
  - Hard disk drive accesses are measured in ms (or in other words, millions of memory accesses)
  - If we scale up, say 1 ns = 1 s, then a single hard disk access takes at least **11.5 days!**

Thinking Beyond Scale

- The value of big datasets lies in the insights they contain
  - How do we extract insights?
    - Queries, iterative refinement
    - Machine learning models
    - Visualizations
- These insights must be timely to be useful
  - Weather predictions
  - Disease spread
  - Sales forecasts (eclipse glasses)

Extracting Insights

- At a basic level, queries allow us to explore relationships between entities in the dataset
  - `SELECT Name FROM Students WHERE Course = 'CS686' AND Current_Location != 'USF'`
- Visualizations can make interactions between **features** in the dataset more obvious
- Machine Learning allows us to **predict** and **classify** large datasets
  - In the past, many of these models just didn't have enough *training samples* to adequately capture the subtleties of the dataset

Wrapping Up

- Big Data is all about:
  1. **Scale**: accomplishing tasks that are just too large/intensive to do on one or a few machines
  2. **Insight**: extracting knowledge from the data (or in business terminology, "extracting value")

Today's Agenda

- Q&A from the previous class
- Defining big data
- **Dataset sources**
- Distributed network design

Sources of Big Data

- Traditional
  - Big files, large quantities, archives, documents
- Sensors
  - Smart devices, radars, satellites, IoT
- The WWW and social media
  - Web crawlers
  - Twitter, Facebook

Data Sources: Traditional

- Movies and photos are being captured at higher resolutions, requiring more storage space
  - Better compression algorithms can mitigate this to some extent, but more people are producing digital media
- Screens are getting better: pixel density on phones and laptops has increased
  - Requires high-resolution graphics/assets
- Using the cloud for storage has become seamless and ubiquitous
  - Ultimately results in more copies of data everywhere

Data Sources: Logs

- Application/OS logs are frequently one of the biggest data sources at organizations
  - Facebook stores 25 TB of logs **per day**
- Logging is vital for:
  - Security – tracing an intrusion
  - Debugging – determining when and where problems occur
  - History – maintaining a record of system activities

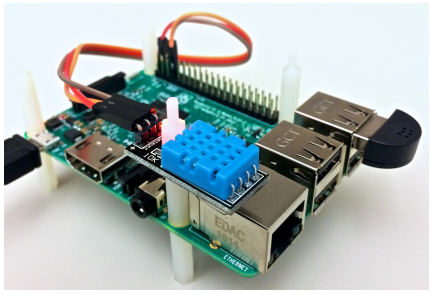
Data Sources: Sensors

- Miniaturization and Internet availability have led to a boom in sensing devices
- We constantly record information about our world
- Why throw this data away when it's so easy to store long-term?
  - More importantly, what can we learn?

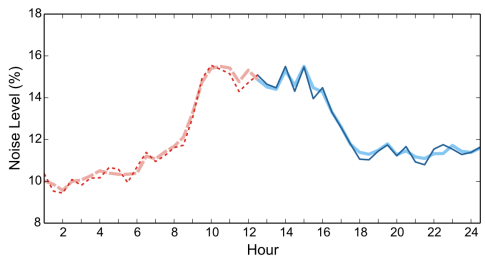
Sensing Devices & Techniques

- Weather radars, satellites, and fixed-location observational devices
- Geolocation (GPS)
  - Where your bus was 30 seconds ago
- Live health monitoring, body measurements, activity tracking
- Click stream data, app usage data
- Autonomous vehicles
- … And your smartphone

IoT Devices: Raspberry Pi

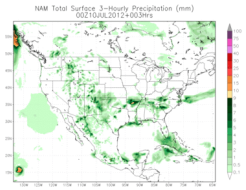


RPi: Monitoring Audio Levels



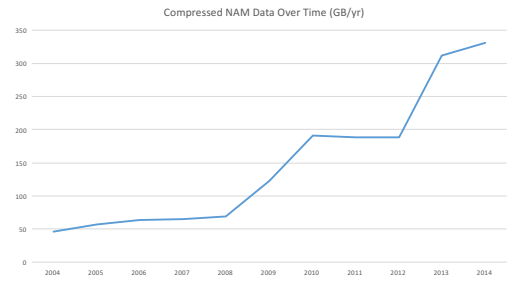
Climate Modeling

- NOAA maintains several climate models to predict and analyze the weather
- This model, the NAM, includes precipitation, humidity, etc.
  - 3-hour intervals

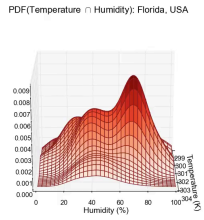


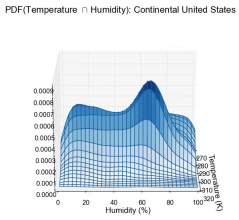
https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/north-american-mesoscale-forecast-system-nam

Increasing Resolution



Climate Analysis





Data Sources: The WWW

- Web search engines were some of the very first big data platforms
  - Index the entire WWW: **web crawler**
  - If you can monitor changes in the web, you can provide better search results
- The Internet Archive stores not only the current state of the web, but also its history
- Most websites implement some level of tracking

Data Sources: Social Media

- Twitter gives researchers access to raw tweet streams through their API
  - Facilitates sentiment analysis
- Social networks like Facebook form large **graphs**
  - We can learn a lot from someone based on their friends, who they follow, and who follows them
- Many platforms can be **scraped** for data

Inspiration

- A great collection of datasets is available at Academic Torrents: http://academictorrents.com
- Includes a variety of sources:
  - Reddit, gaming, stock markets, movie recommendations
- …And a lot of different formats / data types

Data Fusion

- One last concept to think about is **data fusion**
- Some experts claim that when it gets very hot, crime increases
  - Does this have to do with the temperature or is something else at play?
- We can **fuse** two datasets based on time, space or other common features
  - Ties back into the **V**ariety of big data

Today's Agenda

- Q&A from the previous class
- Defining big data
- Dataset sources
- **Distributed network design**

Organizing Big Data

- We have established that dealing with big data is going to require a lot more power than your laptop
- Google and others pioneered *Warehouse Scale Computing* in the early 2000s
  - Their key insight: buying the most powerful hardware is not necessarily the best move!
- Building large clusters of commodity hardware allows businesses to scale

Warehouse-Scale Computing

- In this model, we fill data centers with commodity hardware with the best **dollar:performance** ratio
- Connect the **nodes** with a reasonably-priced interconnect
- Over time, these systems naturally become heterogeneous
- Even more importantly, the nodes are constantly failing… But that's no big deal.